

Identification of truncated forms of U1 snRNA reveals a novel RNA degradation pathway during snRNP biogenesis

Hideaki Ishikawa^{1,2}, Yuko Nobe^{2,3}, Keiichi Izumikawa^{1,2}, Harunori Yoshikawa^{2,4}, Naoki Miyazawa^{2,4}, Goro Terukina^{2,4}, Natsuki Kurokawa^{1,4}, Masato Taoka^{2,3}, Yoshio Yamauchi³, Hiroshi Nakayama^{2,5}, Toshiaki Isobe^{2,3} and Nobuhiro Takahashi^{1,2,4,*}

¹Department of Applied Biological Science, Graduate School of Agriculture, Tokyo University of Agriculture and Technology, 3-5-8 Saiwai-cho, Fuchu-shi, Tokyo 183-8509, Japan, ²Metabolome Division, Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Agency (JST), Sanbancho 5, Chiyoda-ku, Tokyo 102-0075, Japan, ³Department of Chemistry, Graduate School of Sciences and Engineering, Tokyo Metropolitan University, 1-1 Minamiosawa, Hachioji-shi, Tokyo 192-0397, Japan, ⁴Department of Bioengineering, United Graduate School of Agriculture, Tokyo University of Agriculture and Technology, 3-5-8 Saiwai-cho, Fuchu-shi, Tokyo 183-8509, Japan and ⁵Biomolecular Characterization Team, RIKEN Advanced Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

Received February 9, 2013; Revised November 11, 2013; Accepted November 14, 2013

ABSTRACT

The U1 small nuclear ribonucleoprotein (snRNP) plays pivotal roles in pre-mRNA splicing and in regulating mRNA length and isoform expression; however, the mechanism of U1 snRNA quality control remains undetermined. Here, we describe a novel surveillance pathway for U1 snRNP biogenesis. Mass spectrometry-based RNA analysis showed that a small population of SMN complexes contains truncated forms of U1 snRNA (U1-tfs) lacking the Sm-binding site and stem loop 4 but containing a 7-monomethylguanosine 5' cap and a methylated first adenosine base. U1-tfs form a unique SMN complex, are shunted to processing bodies and have a turnover rate faster than that of mature U1 snRNA. U1-tfs are formed partly from the transcripts of U1 genes and partly from those lacking the 3' box elements or having defective SL4 coding regions. We propose that U1 snRNP biogenesis is under strict quality control: U1 transcripts are surveyed at the 3'-terminal region and U1-tfs are diverted from the normal U1 snRNP biogenesis pathway.

INTRODUCTION

Ribonucleoproteins (RNPs) are a class of RNA–protein complexes that facilitate many cellular processes. One of

the most prominent examples is pre-mRNA splicing, which is driven by the spliceosome. The major spliceosomal components are small nuclear RNPs (snRNPs), each of which consists of an snRNA (U1, U2, U4/U6 or U5), a common heptameric ring of Sm proteins (B/B', D1, D2, D3, E, F and G) assembled around the snRNA's Sm-binding site, and several proteins that are unique to each specific snRNP; for instance, the proteins for U1 snRNP are U1-A, U1-C and U1-70K (1). Assembly of Sm proteins on an snRNA is a key step in snRNP biogenesis that takes place in the cytoplasm shortly after the nuclear export of nascent snRNA precursors (pre-snRNAs). Proper assembly of the Sm proteins, 5' cap hypermethylation and 3' end processing of the snRNAs are prerequisites for the subsequent import of snRNPs into the nucleus (1–4).

The remarkable assembly of the seven Sm proteins on the snRNA (5,6) is carried out by a complex containing SMN, a product of *SMN1* that is mutated in the neuromuscular disease spinal muscular atrophy (7). The SMN complex contains eight proteins: Gemin2 (SIP1), 3 (a DEAD-box RNA helicase), 4, 5 [a tryptophan-aspartic acid (WD)-repeat protein], 6, 7, 8 and Unrip (unr interacting protein) (8,9). Importantly, SMN prevents unproductive associations between Sm proteins and RNAs (10–12). Among the components of the SMN complex, Gemin5 determines the specificity for snRNAs; for U1 snRNA, Gemin5 binds pre-U1 snRNA at both the loop region of stem-loop (SL) 1 and the SL4 region (5) directly on its own *via* its WD-repeat domain (13) and delivers pre-snRNAs to sites of Sm core assembly and processing. On the other hand, Gemin2 binds a pentamer of Sm proteins

*To whom correspondence should be addressed. Tel: +81 042 367 5709; Fax: +81 042 367 5709; Email: ntakahas@cc.tuat.ac.jp

containing SmD1, SmD2, SmE, SmF and SmG (14–16). Gemin2 interacts with all five Sm proteins, and its extended conformation enables it to wrap around the entire crescent-shaped pentamer. This prevents the Sm pentamer from assembling on unintended RNAs (12,17). To allow pre-snRNA binding, the N-terminal region of Gemin2 must be displaced from the Sm pentamer's RNA-binding pocket; the mechanistic details of this process, however, remain unclear. Finally, two additional Sm proteins, SmB/B' and SmD3, associate with the Sm pentamer, presumably through direct interaction with SMN (18–23), in a process involving Gemin3, 4, 6, 7, 8 and Unrip (4,24–28).

In *Drosophila*, the assembly of Sm proteins on U snRNAs takes place in cytoplasmic U-bodies that invariably associate with processing bodies (P-bodies), which function in RNA surveillance and turnover (29). Several lines of evidence suggest that SMN is required for the functional integrity of the U-body–P-body pathway, which is important for maintaining proper nuclear architecture in germline cells (30). However, how the functional integrity of the U-body–P-body pathway is related to snRNP biogenesis and how aberrant snRNPs are discriminated from normal snRNPs during their biogenesis remain unknown.

Here we used a mass spectrometry (MS)-based method to directly identify cellular RNAs in an unbiased manner. This high-throughput approach has previously yielded information regarding RNA modifications (31–34), and its use allowed us to identify novel truncated forms of U1 snRNA (U1-tfs) that have a 5'-monomethylguanosine (m⁷G) cap and base methylation at the first transcribed nucleotide but lack the Sm-binding site and SL4 region. U1-tfs are generated *de novo*; form complexes with the phosphorylated adaptor for nuclear export (PHAX), SMN, Gemin2, 3, 4, 5, 6 and 8; and are mostly localized in P-bodies. We show that failure of Sm protein loading on U1-tfs is responsible for the previously observed localization of the SMN complex in P-bodies (29,30).

MATERIALS AND METHODS

MS-based RNA analysis

RNAs were separated on a Develosil C30-UG-2 column (3 μ m, 2.0 mm i.d., 100-mm long, Nomura Chemical) at 60°C by gradient elution at a flow rate of 50 μ l/min. The solvents were A, 400 mM hexafluoroisopropanol/triethylamine (pH 7.0) and B, 100% methanol. After applying an RNA mixture, the column was eluted with 10% B, followed by a gradient to 23% B over 1 min and then to 28% B over 60 min. RNAs were detected by monitoring A₂₆₀ and, where necessary, the fractionated RNA was digested with RNase T1 (10 ng) in 10 mM sodium acetate (pH 5.3) at 37°C for 60 min. The resulting oligonucleotide mixture was then analyzed by the nano-flow LC-tandem (MS/MS) system (32,33) equipped with a genome-oriented database search engine Ariadne (31). We used the genome database of *Homo sapiens* (reference assembly version GRCh37 obtained from ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/) under the following

parameters: maximum number of missed cleavages, 0; variable modification parameter, one methylation per RNA fragment for any residue; RNA mass tolerance, ± 20 ppm; and MS/MS tolerance, ± 750 ppm.

Cloning and construction of plasmids for exogenous expression of U1 snRNA or its truncated mutants

To generate plasmids to exogenously express U1 snRNA, we first amplified the region encoding human U1 (chromosome 1, gene ID 26871) and flanking regions from human genomic DNA extracted from HEK293 cells for use as the PCR template with the primer set 5'-GAAGGATCCGTTTCTTTTGTAAATCCGAAACA-3' and 5'-CAACTCGAGCTCTATGAGGTGAGAACACACT-3'. The amplified DNA fragment was digested with BamH I/Xho I and ligated into the corresponding sites of pcDNA3.1. After verifying the sequence of the U1 gene-containing DNA fragment, it was excised with BamH I/Xho I and ligated into pcDNA3.1 (pcDdCMV-U1; WT); the cytomegalovirus (CMV) promoter was removed with Bgl II/Xho I, thereby avoiding CMV promoter-based transcription.

To construct U1 snRNA-expressing plasmids lacking the Sm or Sm-SL4 region or the cis-acting DNA elements DSE, PSE or 3'box (Δ Sm, Δ SmSL4, Δ DSE, Δ PSE, Δ 3'box), PCR was done using WT vector as template and the primer set 5'-TATGCAGTCGAGTTTCCCACATTTG-3' and 5'-TAGTGGGGGACTGCGTTTG-3' for Δ sm, 5'-TATGCAGTCGAGTTTCCCACATTTG-3' and 5'-ACTTTCTGGAGTTTCAAAAACAGAC-3' for Δ smSL4, 5'-CTGTCCGTGATGTCACCGACAG-3' and 5'-GGCAGCGCAGAGGCTGCTG-3' for Δ DSE, 5'-GCCCCGCGCACTCCCGAG-3' and 5'-GAGTGAGGCGTATGAGGCTGTGTC-3' for Δ PSE or 5'-TCCAGAAAGTCAGGGGAAAGC-3' and 5'-CCGTACGCCAAGGGTCATGTC-3' for Δ 3'box. The PCR products having blunt ends were phosphorylated by T4 polynucleotide kinase and then self-ligated. Deletion of a specific sequence element in each construct was verified by DNA sequencing.

To generate the RNA tag-fused U1 snRNA expression vector, the U1 coding sequence was PCR amplified from WT U1 template using the primer set 5'-GGAATCGATA TACTTACCTGGCAGGGGAG-3' and 5'-CAACTCGA GCTCTATGAGGTGAGAACACACT-3' to add a Cla I site near the 5' end of U1 snRNA. The amplified DNA fragment encoding U1 snRNA was cleaved with Cla I/Xho I.

The y18Sn tag was prepared using the oligonucleotides 5'-GGAAGATCTCATACTTACCTGCGAGGATTTCAG GCTTTGGATCGATGGA-3' and 5'-TCCATCGATCC AAAGCCTGAATCCTCGCAGGTAAGTATGAGATC TTCC-3, which were annealed and cleaved with Bgl II/Cla I. In the former primer sequence, the first 11 nucleotides of U1 snRNA (ATACTTACCTG) were included just before the 5' end of the aptamer tag sequence. We used the Bgl II site present in the region just before the U1 coding sequence within the U1 gene sequence for construction. This was done based on a report that at least the first nine nucleotides of U1 snRNA are required for recognition of canonical 5' pre-mRNA splice site (35).

On the other hand, RAT tag (36) was synthesized using the primer set 5'-GGAAGATCTCATACTTACCTGTAA GGAGTTTATATGGAAACCCCTTAGACGTCGGCAC GAGGTTTAG-3' and 5'-TCCATCGATGGCAGGAGT GTAGCTAAACCTCGTGCCGACGTC-3', which were annealed and amplified by PCR. The amplified DNA fragment was cut with Bgl II/Cla I. The DNA fragments encoding U1 snRNA prepared with Cla I/Xho I and that encoding y18Sn or RAT tag prepared with Bgl II/Cla I were ligated into the Bgl II/Xho I site of WT vector. The vector expressing y18Sn-U1 snRNA was, named, y18Sn-WT, and that expressing RAT-U1 snRNA was named RAT-WT. All constructs were verified by DNA sequencing.

RNA aptamer-based affinity purification

To isolate RNPs, we used the RAT tag-based method (36) with modifications. Briefly, instead of using purified recombinant PP7 coat protein (PP7CP), we cotransfected two plasmids (total 30 µg) into 293T cells (5×10^7 in DMEM) using the calcium phosphate method; one plasmid expressed RAT-tagged RNA and another expressed HA-FLAG (HF)-tagged PP7CP (pcDNA3.1-PP7CP-HF) in a 7:1 ratio. At 24 h post-transfection, cells were collected and rinsed with phosphate-buffered saline-calcium/magnesium-free [PBS(-)] once, lysed in P100 lysis buffer [50 mM Tris-HCl (pH 8.0), 100 mM KCl, 5 mM MgCl₂, 0.5% IGEPAL-CA630, 1 mM PMSF] and incubated on ice for 30 min. The soluble fraction, prepared by centrifugation at 20 000g for 30 min at 4°C, was mixed with anti-FLAG-M2-conjugated agarose beads and rotated for 3 h at 4°C. PP7CP-HF bound to RAT-tagged RNA-protein complexes was captured by the beads. After washing five times with P100 lysis buffer and once with P100 lysis buffer without IGEPAL-CA630, RAT-tagged RNA-protein complexes were eluted from the beads with 500 µg/ml FLAG peptide in the same buffer.

Fluorescence *in situ* hybridization

Fluorescence *in situ* hybridization (FISH) was carried out as described by Rouquette *et al.* (37). Briefly, HeLa or 293T cells were seeded on collagen-coated culture slides, fixed with 4% paraformaldehyde in PBS, washed twice with PBS and permeabilized for 16 h at 4°C in 70% ethanol. After washing two times with 2× SSC containing 10% formamide to rehydrate the cells, hybridization was done with 0.5 ng/ml of a fluorescently labeled DNA probe in hybridization solution (10% formamide, 2× SSC, 0.5 mg/ml yeast tRNA, 10% dextran sulfate, 50 µg/ml BSA, and 10 mM ribonucleoside vanadyl complexes) for 3 h at 37°C. The following fluorochrome-labeled DNA probes were used: 5'-Cy3-labeled y18Sn probe (5'-CCAAAGCCTGAATCCTCG-3'), 5'-FITC-labeled U1-#1 probe (5'-GTATCTCCCCTGCCAGGTAAGTA T-3'), 5'-FITC-labeled U1-#3 probe (5'-TATGCAGTCG AGTTTCCCACATTTGG-3'), 5'-FITC-labeled U1-#SL4 probe (5'-GAAAGCGCGAACGCAGTCCCCCACTA-3'), 5'-Cy3-labeled U2 probe (5'-CTACACTTGATCTTA GCCAAAAGGCCGAGAAGC-3') and 5'-Cy3-labeled

5'ITS1 probe (5'-CCTCGCCCTCCGGGCTCCGTAA TGATC-3'). When a mixture of Cy3- and FITC-labeled probes was used, the concentration of each probe was 0.5 ng/ml. After hybridization, the cells were washed twice with 2× SSC, 10% formamide and once with PBS. Immunocytochemistry was carried out as described in Supplementary information. The staining was observed with an Axiovert 200 M microscope.

RESULTS

MS-based analysis of SMN-associated RNAs identifies U1-tfs

We first expressed WT SMN from a single-copy transgene encoding a triple affinity-purification (DAP) tag (biotin and FLAG tags useful for visualization and purification and an additional N-terminal 6-histidine epitope tag; Figure 1A) (38). The expressed DAP-SMN localized not only in the cytoplasm but also in the nucleus as foci, as reported for endogenous SMN (Supplementary Figure S1A). Pull-down using DAP-SMN as bait showed that DAP-SMN associated with Gemins 2, 3, 4, 5, 6 and 8, Unrip, coilin, SmB/B', SmD1, SmE and U1A as reported (11,39,40) (Supplementary Figure S1B). Northern blotting showed that the DAP-SMN complexes contained, minimally, U1, U2, U4, U5, U7 and U11 snRNAs, each of which is known to associate with SMN (12,39) (Figure 1B), suggesting that DAP-SMN was functionally indistinguishable from endogenous SMN.

SYBR Gold-stained RNA bands were excised, digested with RNase T1 and analyzed by nanoflow LC-MS/MS in combination with Ariadne, a genome-wide search engine for RNA identification (31). We also analyzed the pulled-down RNAs by the LC method, where RNAs were separated by reverse-phase semimicro-LC, fractionated, digested with RNase T1 and analyzed by nanoflow LC-MS/MS and Ariadne (Figure 1C and Supplementary Tables S1 and S2). These analyses identified, minimally, U1, U2, several forms of U5, U6, U7 and U11 snRNAs (Figure 1B and C and Supplementary Tables S1 and S2). In addition to U1 snRNA, LC-MS/MS identified U1-tfs in a stained band (~120 nt) that migrated faster than mature U1 snRNA (Figure 1B) and in the earlier eluting fractions relative to mature U1 snRNA (Figure 1C). On RNase T1 digestion, this U1-tfs generated 21 distinct oligonucleotides originating from the region 1–121 of mature U1 snRNA (Supplementary Table S2), but we did not detect oligonucleotides from the 3' region 122–164. To confirm the MS-based identification of U1-tfs, we used four DNA probes complementary to regions 1–24 (probe #1), 65–86 (#2), 100–124 (#3) and 125–147 (#4) of mature U1 snRNA (Figure 1D) and performed northern blotting to detect the RNAs extracted from DAP-SMN complexes or total cellular RNA. U1-tfs were detected by probes #1, #2 and #3 but not #4 (Figure 1D); these results agreed with those of the MS-based identification. We also analyzed the nucleotide sequences of RNAs extracted from the SYBR Gold-stained U1-tfs band (Figure 1B) by the 3' rapid amplification of

cDNA ends protocol, which confirmed that each sequence analyzed had a length of 118–127 nt starting from the 5' end of the mature U1 snRNA (Supplementary Table S3). These results clearly indicated

that U1-tfs lack the Sm-binding site and SL4 region. In addition, the results for total RNA indicated that U1-tfs were not an artifact formed during the pull-down of DAP-SMN complexes (Figure 1D).

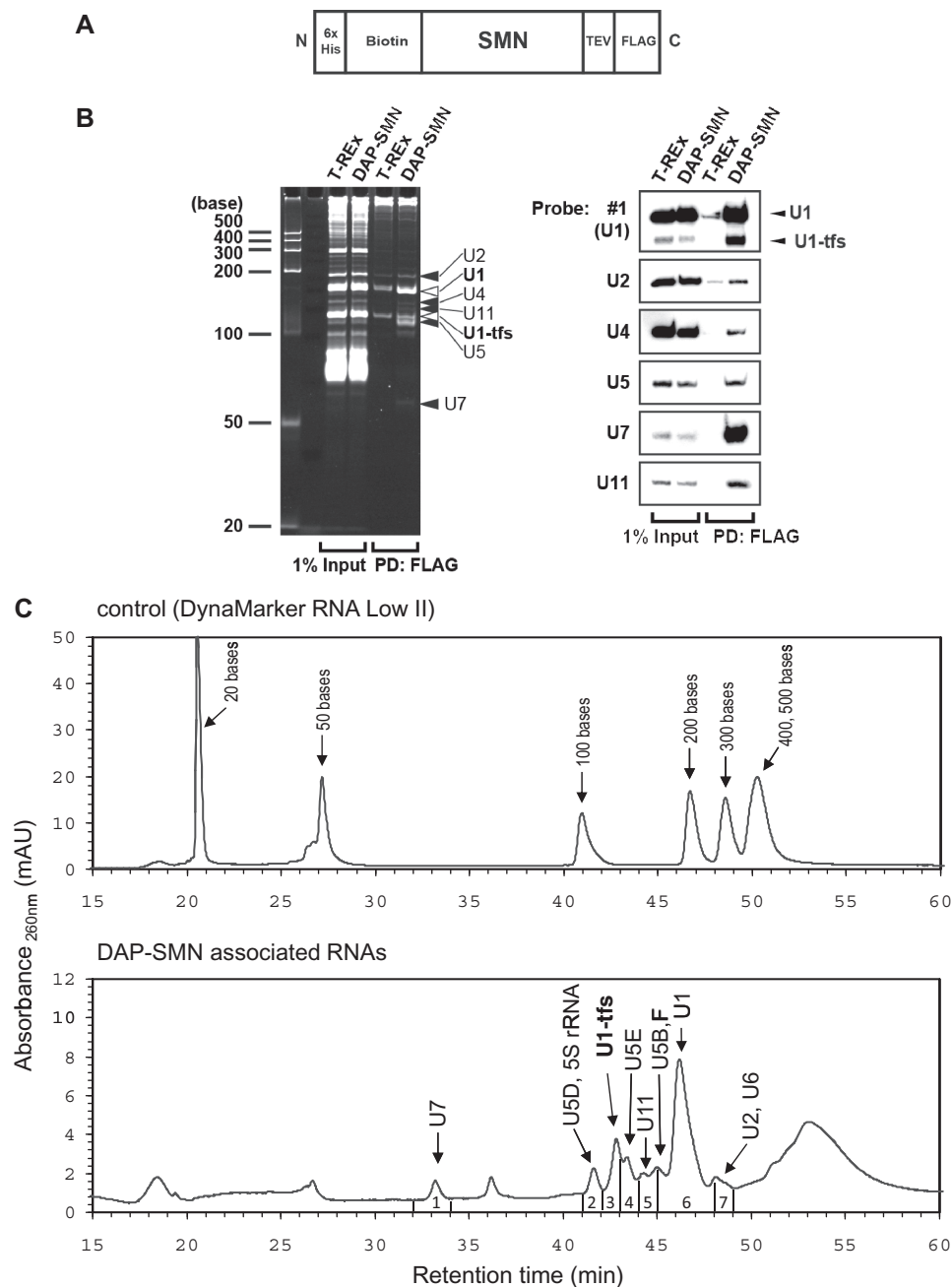


Figure 1. Identification of U1-tfs as a novel RNA component associated with SMN. (A) Schematic diagram of DAP-tagged SMN (DAP-SMN) used for pull-down analysis. SMN was fused at the N-terminal (N) with a 6-histidine tag and biotinylation sequence and at the C-terminal (C) with a FLAG tag connected by a TEV cleavage sequence. (B) DAP-SMN-associated RNAs were separated by denaturing PAGE (7.5 M urea, 9% polyacrylamide). RNAs were visualized by SYBR Gold staining (left) or by northern blotting with probes complementary to RNA as indicated (right). Input is the cell extract used for pull-down analysis. DAP-SMN: extract of DAP-SMN-expressing T-REx 293 cells. T-REx: extract of control parent cells, PD: pull-down. (C) A typical chromatogram of standard RNAs (upper) or of the DAP-SMN-associated RNAs (bottom) separated by reverse-phase LC. Each fraction (1–7) was subjected to LC-MS/MS-Ariadne analysis. Sizes of the standard RNAs or the specific RNAs identified by MS are indicated. (D) Northern blot analysis for DAP-SMN-associated RNA or total RNA of HeLa cells. The analysis was done with the probes (#1, #2, #3 and #4) corresponding to the regions that are shown under the sequence of U1 snRNA (top). The region corresponding to #SL4 probe is also given. DAP-SMN-associated RNA or total RNA was separated by denaturing urea-PAGE and stained with SYBR Gold or analyzed by northern blotting with probes #1–4.

(continued)

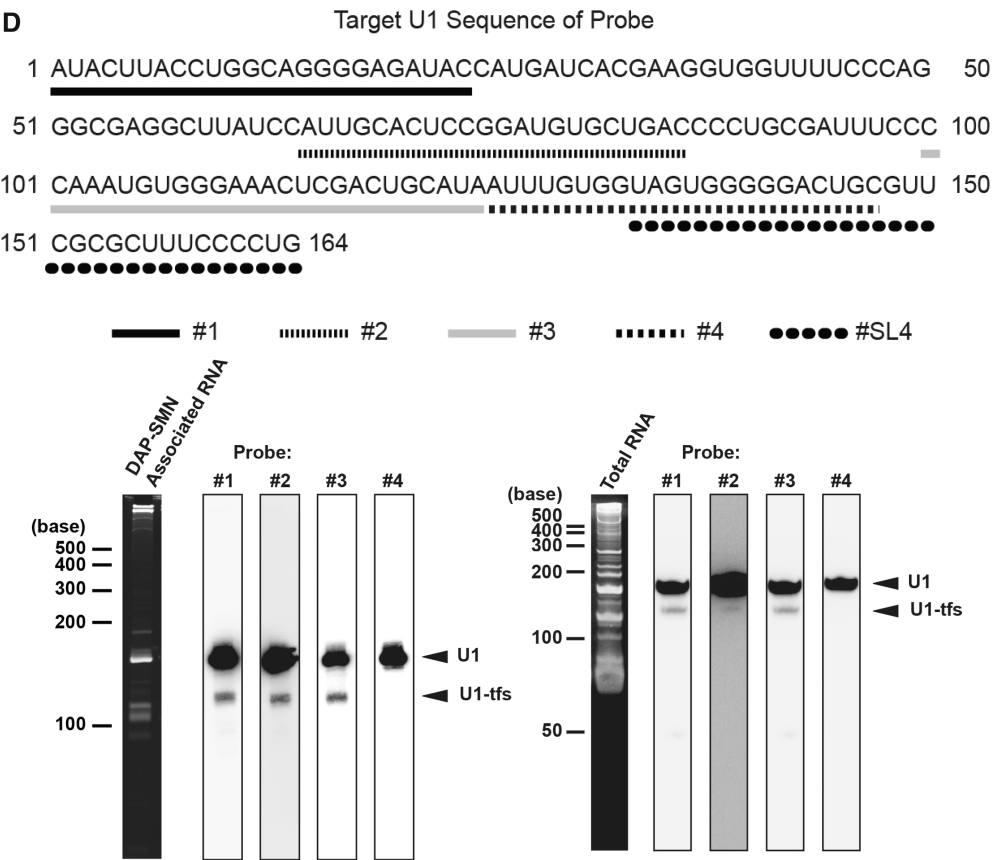


Figure 1. Continued.

U1-tfs have a 7-monomethylguanosine 5' cap and a methylated first adenosine base and lack the methylation at position 70

During the LC-MS/MS analyses, collision-induced dissociation generated a series of product ions from the RNA fragments. The product ions included the major c/y and a/w series with minor derivatives (dehydrated ions and ions that lost nucleotide bases; Supplementary Figure S2A and Supplementary Tables S1 and S2), as reported (33). Among the oligonucleotides generated from the U1 snRNA fractions (Figure 1C), we mainly detected m₃GpppAmUmACUUACCUGp (or m₃GpppAmUmACΨΨACCUGp) (containing a 5' cap of 2,2,7-trimethylguanosine, m₃G) that originated from the 5' end (Figures 2A, Supplementary Figure S2B and C and Supplementary Table S2) and CUUUCCCCUG-OH from the 3' end (Supplementary Figure S2D and Supplementary Table S2), indicating that the oligonucleotides mostly originated from mature U1 snRNA. We also detected CUUUCCCCUG>p (2', 3' cyclic phosphate) in a ratio of ~1:6 with CUUUCCCCUG-OH (estimated using relative mass intensities; Supplementary Figure S2D and Supplementary Table S2). Because RNase T1 cleavage should generate CUUUCCCCUG>p, those SMN-associated RNAs probably contained at least ≥1 extended bases at the 3' end. The MS-based method could distinguish a one-nucleotide extension as well as

the absence of a phosphate or hydroxyl group at the 3' terminus of the mature U1 snRNA. In addition, we detected methylation at position 70 in ~90% of the oligonucleotides corresponding to the region 69–75 of mature U1 snRNA (CAmCUCCG>p; Supplementary Table S2, Figure 2B and Supplementary Figure S2E), as reported (41). On the other hand, one oligonucleotide generated from the 5' end region of U1-tfs had a mG cap (Figure 2A) and three methyl groups in the first two nucleotides (Supplementary Figure S2B and C); two of the latter reflected ribose 2' O-methylation at the first and the second transcribed nucleotides, and the remaining methyl group was found in the adenine base of the first nucleotide (m⁷GpppmAmUmACp; Supplementary Table S2 and Supplementary Figure S2B and C). In addition, U1-tfs produced almost exclusively an oligonucleotide without the methylation at position 70 (CACUCCG, 69–75; Supplementary Table S2, Figure 2B and Supplementary Figure S2E). Supplementary Figure S2F summarizes the secondary structure of U1-tfs identified by MS-based analysis in comparison with mature U1 snRNA.

U1-tfs associate with proteins involved in early steps of U1 snRNP biogenesis in the cytoplasm

Given that U1-tfs have an m⁷G cap, we examined whether U1-tfs associate specifically with proteins involved in early steps of U1 snRNP biogenesis. We used PHAX,

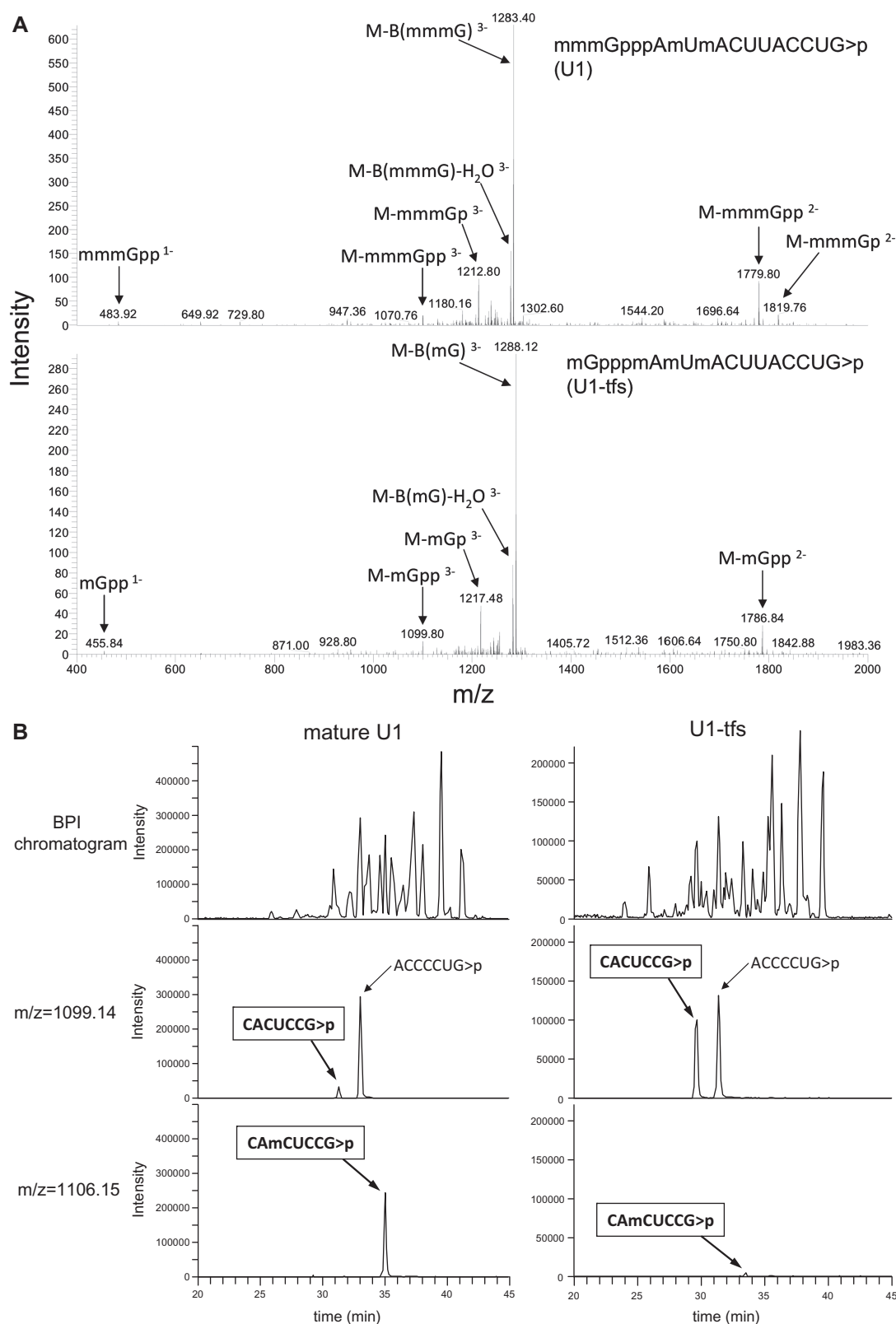


Figure 2. Assignment of post-transcriptional modifications of U1-tfs using LC-MS/MS. **(A)** MS/MS spectrum of the 5'-terminal oligonucleotide derived from RNase T1 digest of U1 snRNA (upper left) or U1-tfs (bottom left). Various MS/MS fragments originating from the m₃G cap of U1 snRNA or from the m⁷G cap of U1-tfs are indicated. Nomenclature for the product ions generated by collision-induced dissociation is shown in Supplementary Figure S2A. **(B)** Reverse-phase LC separation of the RNase T1 digest of U1 snRNA or U1-tfs. Effluent was monitored as the count of total ions (top), $m/z = 1099.14$ (CACUCCG>p; middle) or $m/z = 1106.15$ (CAmCUCCG>p; bottom). ACCCCUG>p is isobaric with CACUCCG>p and is present in both RNA digests. BPI chromatogram: chromatogram monitored by BPI [BPI: Base Peak Intensity; the most intense (highest signal) peak in a mass spectrum in the range of m/z values measured].

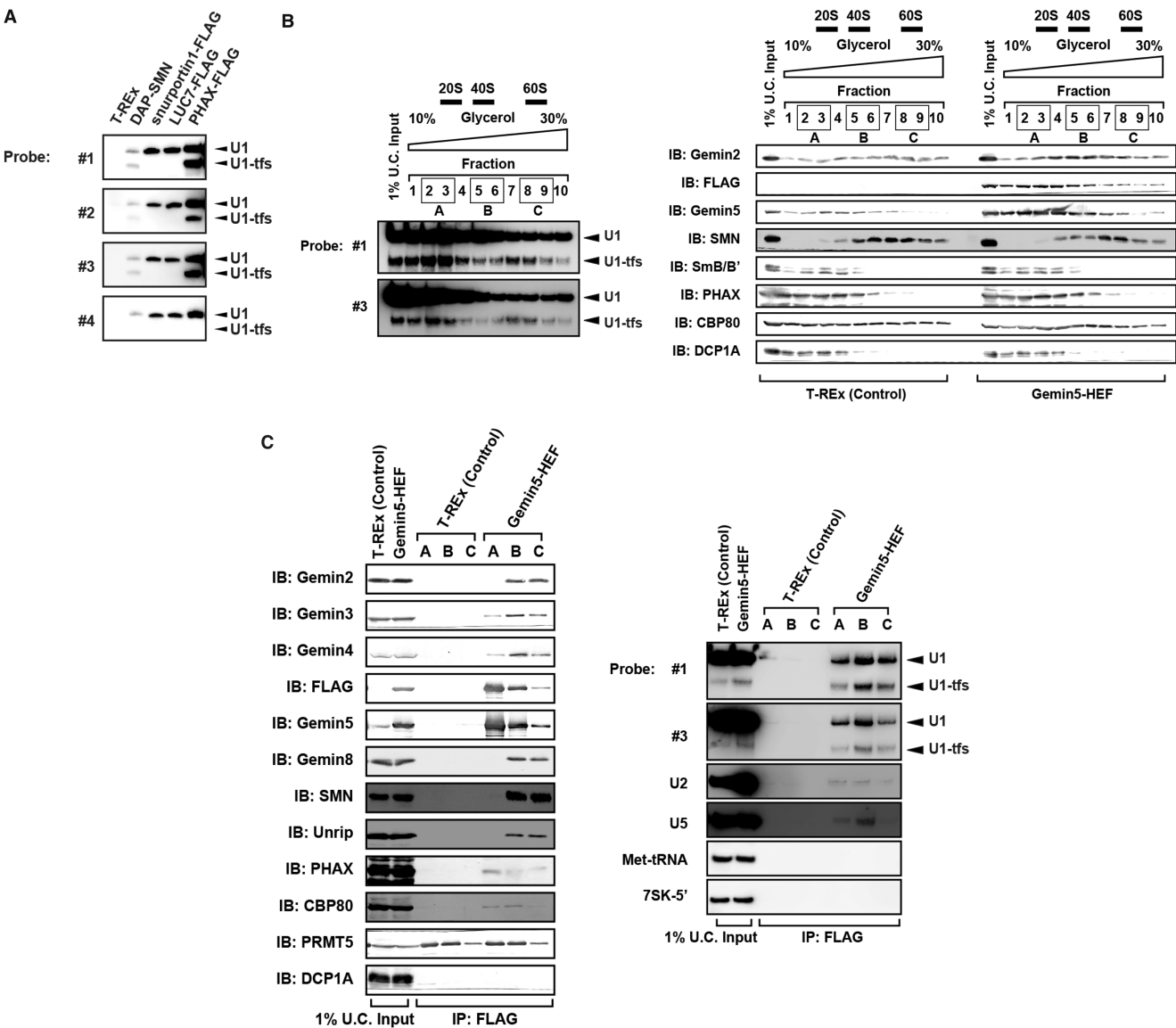


Figure 3. U1-tfs are formed at early steps of U1 snRNP biogenesis. (A) RNAs were pulled down using DAP-SMN, FLAG-fused PHAX, U1-70K, snurportin or LUC7 as affinity bait. RNAs were separated by denaturing urea-PAGE and visualized by northern blotting with probes #1–4. (B) Total cell extract of T-REx 293 cells (left panel) or Gemin5-HEF-expressing T-REx 293 cells (right panel) were separated into 10 fractions by glycerol gradient (10–30%) ultracentrifugation. Each fraction was analyzed by immunoblotting with antibodies against the proteins indicated. Northern blotting with DNA probe #1 or #3 showed a typical U1 RNA distribution pattern of this ultracentrifugation fractionation. Sedimentation values are indicated above the elution profile. (C) Fractions 2 and 3 (mixture A), 5 and 6 (mixture B) or 8 and 9 (mixture C) in (B) were mixed and subjected to pull-down analysis with Gemin5-HEF as affinity bait. Proteins and RNAs pulled down with Gemin5-HEF were analyzed by immunoblotting (IB) with antibodies against the proteins indicated and by northern blotting with the DNA probes indicated, respectively.

snurportin and LUC7 as affinity bait to pull down U1-tfs and constructed doxycycline-inducible cell lines for their expression. PHAX associates with pre-snRNAs soon after their transcription, and the PHAX–pre-snRNA complex is exported with CBP20/80 and other proteins to the cytoplasm (42). Snurportin recognizes the m³G cap of U1 snRNA in the cytoplasm and is imported with importins into the nucleus. LUC7 is a human homolog of budding yeast Luc7p, which recognizes and binds the 5' splice site of pre-mRNA (43). Pull-down analyses showed that U1-tfs associated with PHAX but not snurportin or LUC7 (Figure 3A). This result was consistent with the fact that the U1-tfs cap is only monomethylated and suggested that

U1-tfs associate with at least SMN and PHAX and are only present in early steps of snRNP biogenesis, i.e. before 5' cap hypermethylation.

To further analyze the SMN complexes associated with U1-tfs, we used glycerol gradient ultracentrifugation to prepare 10 fractions of the total cell extract of T-REx 293 cells or doxycycline-inducible Gemin5-hemagglutinin/TEV protease cleavage site/FLAG (HEF)-expressing T-REx 293 cells. U1-tfs eluted throughout the fractions with two distinct maximum peaks at fractions 3 and 7, whereas SMN eluted in fractions 3–10 with a single peak at fraction 7 (Figure 3B). We mixed fractions 2 and 3 (mixture A), 5 and 6 (mixture B) and 8 and 9 (mixture

C), pulled down the protein in each mixture using Gemin5-HEF as affinity bait and identified the pulled-down proteins by immunoblot analysis. This analysis showed that Gemin5-HEF associated with CBP80, PHAX, Gemin3 and Gemin4 as well as with mature U1 snRNA and U1-tfs in mixture A (Figure 3C). However, SMN, Gemin2, Gemin8 and Unrip were not detected in this Gemin-HEF-associated complex (Figure 3C). On the other hand, in mixtures B and C, Gemin5-HEF associated with Gemin2, Gemin3, Gemin4, Gemin8, Unrip and SMN as well as with mature U1 snRNA and U1-tfs (Figure 3C). We further examined the association of U1-tfs with Gemin2, Gemin6, SmB/B', SmD1 or SmE by using its HEF-fused form expressed in the corresponding doxycycline-inducible cell line as the affinity bait and showed that all of those proteins were associated with U1-tfs (Supplementary Figure S3). These results suggested that U1-tfs formed at least two distinct complexes—one containing Gemin3, Gemin4, PHAX and Gemin5 but lacking Gemin2, Gemin8, Unrip, and SMN, and the other containing many known components of SMN complexes including Gemin2, Gemin8, Unrip and SMN. Thus, U1-tfs could be formed at early stages of U1 snRNP biogenesis at least before the association with SMN.

U1-tfs are formed from transcripts of the WT U1 gene construct and that lacking the 3' box element or having defects in the SL4 region

Given the fact that there are a number of *bona fide* loci for U1 snRNA gene and pseudogene in human genome, we examined a possibility that U1-tfs arose from some of those *bona fide* loci. The human genome has three U1 genes containing a 164-base U1-coding region (SL1, SL2, SL3, Sm, SL4) and three cis-acting elements—distal sequence element (DSE), proximal sequence element (PSE) and 3'box (35,44). We used vector pcDNA3.1 to construct various U1 snRNA genes, namely, WT, Δ SmSL4, Δ DSE, Δ PSE and Δ 3'box (Supplementary Figure S4A and B). WT vector contained all of the elements and the region containing the 380-nt sequences downstream of the 3'box reported in the U1 gene (ID 26871) and was used as a control for the full gene expressing mature U1 snRNA (Supplementary Figure S4A and B). Δ SmSL4 lacked the Sm-binding site and SL4 region and was used to express U1-tfs. Δ DSE, Δ PSE and Δ 3'box lacked the DSE, PSE and 3'box region, respectively (Supplementary Figure S4B). We transiently transfected each vector into 293 EBNA cells [human embryonic kidney cell line that stably express Epstein Barr Virus (EBV) EBNA-1 gene from pCMV/EBNA] and detected U1-tfs in total cellular RNA by northern blotting with probes #1 and #3. U1-tfs level increased significantly on transfection with Δ SmSL4 or Δ 3'box but not with WT, Δ DSE or Δ PSE (Supplementary Figure S4C). We also constructed vectors containing U1 genes fused with an RNA tag that distinguishes exogenously expressed U1 snRNA from endogenous U1 (Figure 4A). For the RNA tag, we used y18Sn (yeast 18S neutral), which does not form a stable secondary structure and is expected not to inhibit normal RNA activity (45), or RAT (RNA

Affinity in Tandem), which was previously used to affinity purify 7SK RNP (36). We first constructed an additional six expression vectors (y18Sn-WT, 199 nt; y18Sn- Δ SmSL4, 160-nt; y18Sn- Δ DSE, 199 nt; y18Sn- Δ PSE, 199 nt; y18Sn- Δ 3'box, 199 nt; y18Sn- Δ Sm, 191 nt), each of which had an extra ATACTTACCTG sequence and an y18Sn at the 5' end of the U1 coding region (Figure 4A). The ATACTTACCTG sequence corresponds to the first 11 nucleotides from the 5' end of U1 snRNA, and its presence at the 5' terminus is required for base pairing with a canonical 5' splice site (35). Transient expression of those vectors and northern blotting with a probe complementary to y18Sn confirmed the results obtained for expression of WT, Δ SmSL4, Δ DSE, Δ PSE or Δ 3'box; namely, absence of the 3'box increased the level of U1-tfs (Figure 4B). These analyses showed that y18Sn-WT also formed U1-tfs (Figure 4B). In addition, these analyses revealed that the y18Sn- Δ DSE produced a full-length mature U1 snRNA but not U1-tfs, whereas y18Sn- Δ PSE did not produce a detectable transcript (Figure 4B). Interestingly, y18Sn- Δ Sm expression produced mostly the transcript of expected size and produced U1-tfs at a ratio similar to that of y18Sn-WT (Figure 4B). As shown in Supplementary Figure S4D, we obtained similar results using RAT-tagged expression vectors, indicating that the formation of U1-tfs did not depend on the tag used. Overall these results suggested that both WT and Δ 3'box U1 genes contribute to the formation of U1-tfs, whereas Δ 3'box U1 gene appears to produce U1-tfs more efficiently than WT gene.

We next examined a possibility that U1-tfs arose from some structural defects in U1 snRNA transcripts. We constructed six other variant expression vectors (y18Sn- Δ SL4-1, y18Sn- Δ SL4-2, y18Sn- Δ SL4-3, y18Sn- Δ SL4-1 Δ 3'box, y18Sn- Δ SL4-2 Δ 3'box, y18Sn- Δ SL4-3 Δ 3'box; Figure 4C). y18Sn- Δ SL4-1 lacked the SL4 region and was expected to produce a 172-nt y18Sn-U1 snRNA, whereas y18Sn- Δ SL4-2 (185 nt) and y18Sn- Δ SL4-3 (186 nt) lacked the first and last half of SL4, respectively. The other three constructs, y18Sn- Δ SL4-1 Δ 3'box, y18Sn- Δ SL4-2 Δ 3'box and y18Sn- Δ SL4-3 Δ 3'box, lacked the 3'box region of y18Sn- Δ SL4-1, y18Sn- Δ SL4-2 and y18Sn- Δ SL4-3, respectively. Despite the different sizes of the expected transcripts, transient expression of these six vectors almost exclusively yielded U1-tfs (Figure 4D). As an exception, y18Sn- Δ SL4-1 expression produced U1-tfs as well as an RNA species having the size of U1 snRNA lacking SL4; however, this RNA species was not observed on expression of y18Sn- Δ SL4-1 Δ 3'box (Figure 4D). Use of the RAT-tagged vectors yielded results similar to those obtained for the y18Sn-tag constructs; e.g. defects in the SL4 region primarily yielded U1-tfs (Supplementary Figure S4E). Thus, the formation of U1-tfs is related to the transcriptional events of the gene that lacks the 3'box or has deficiency in the SL4 region.

The U1-tfs snRNA localizes primarily to P-bodies

We used FISH to examine the localization of U1-tfs in cells. We transiently expressed y18Sn-WT, y18Sn- Δ SmSL4 or y18Sn- Δ SL4-1 in 293T cells and detected y18Sn-WT

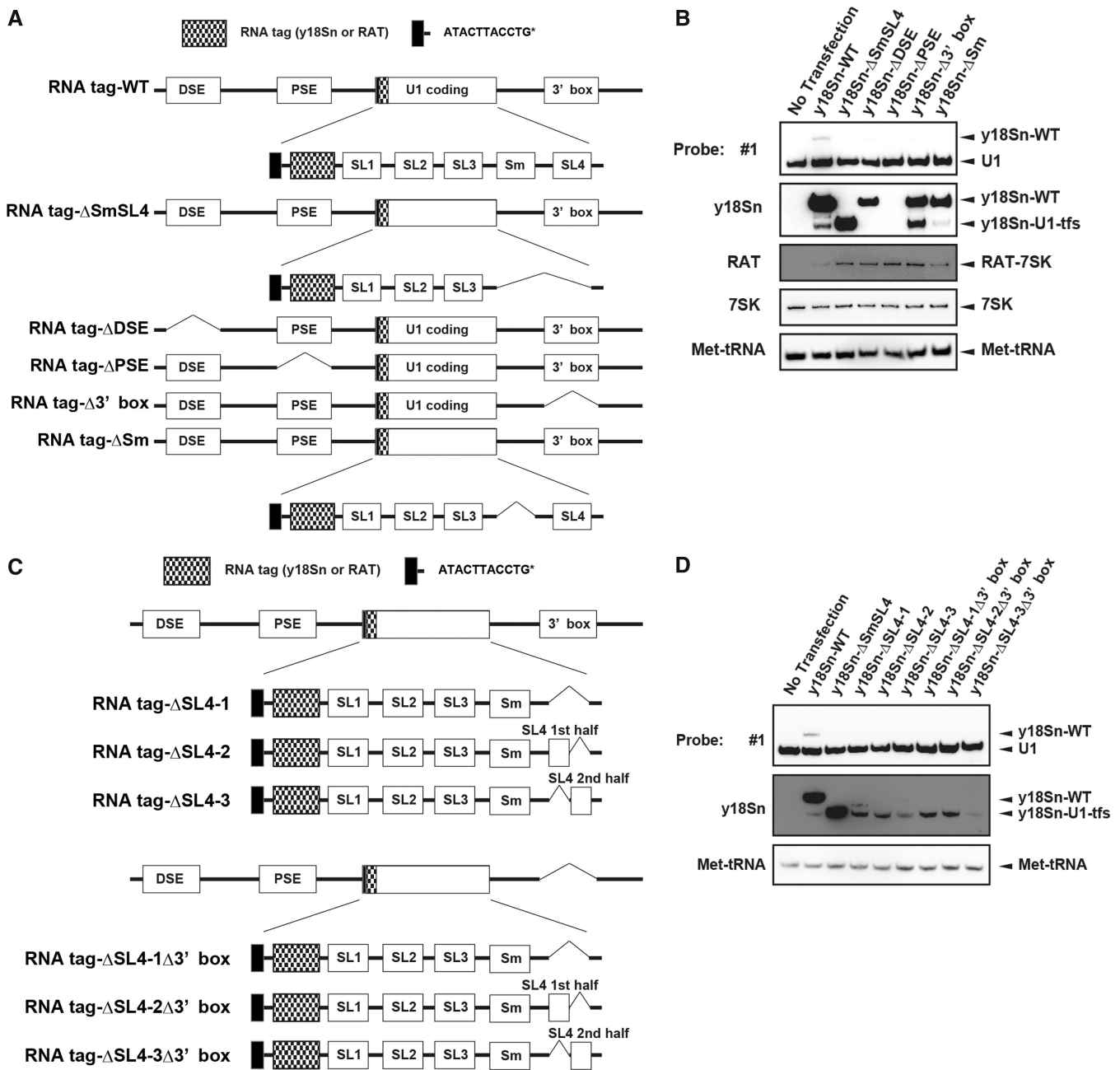


Figure 4. Formation of U1-tfs depends on transcription from the U1 gene. (A) Schematic diagram of the U1 gene construct expressing RNA-tagged (y18Sn or RAT)-U1 snRNA or its deleted forms. See the text for the explanation of each construct (RNA tag-WT, y18Sn-WT or RAT-WT; RNA tag-ΔSmSL4, y18Sn-ΔSmSL4 or RAT-ΔSmSL4; RNA tag-ΔPSE, y18Sn-ΔPSE or RAT-ΔPSE; RNA tag-ΔDSE, y18Sn-ΔDSE or RAT-ΔDSE; RNA tag-Δ3' box, y18Sn-Δ3' box or RAT-Δ3' box; and RNA tag-ΔSm, y18Sn-ΔSm or RAT-ΔSm). (B) Total RNA was extracted from 293 EBNA cells transiently transfected with an expression vector encoding one of the y18Sn-tagged constructs. RNAs were analyzed by northern blotting with a probe complementary to the y18Sn sequence or with probe #1. An expression vector encoding RAT-7SK was used as a transfection control. (C) Schematic diagram of an RNA-tagged (y18Sn or RAT)-U1 gene construct (with or without 3' box) having a defect in the SL4 region. Each construct is explained in the text. (D) Total RNA extracted from 293T cells transiently transfected with an expression vector composed of one of the y18Sn-tagged constructs was analyzed by northern blotting with the y18Sn probe or probe #1.

or y18Sn-U1-tfs using the Cy3-labeled probe for y18Sn tag (Cy3-y18Sn). Endogenous U1 snRNA and exogenously expressed U1 snRNA were also detected using the FITC-labeled probe for region SL1, SL3 or SL4 (FITC-#1, FITC-#3 or FITC-#SL4) corresponding to the sequence of probe #1 or #3 used for northern blotting or #SL4

shown in Figure 1D. FISH using a sequence complementary to y18Sn or U1 snRNA (probe #1) indicated that y18Sn-WT localized almost exclusively in the nucleoplasm (Figure 5A), suggesting that the y18Sn-tag did not interfere with U1 snRNA localization. Cy3-y18Sn staining resulted in a number of cytoplasmic dots in the y18Sn-ΔSmSL4-

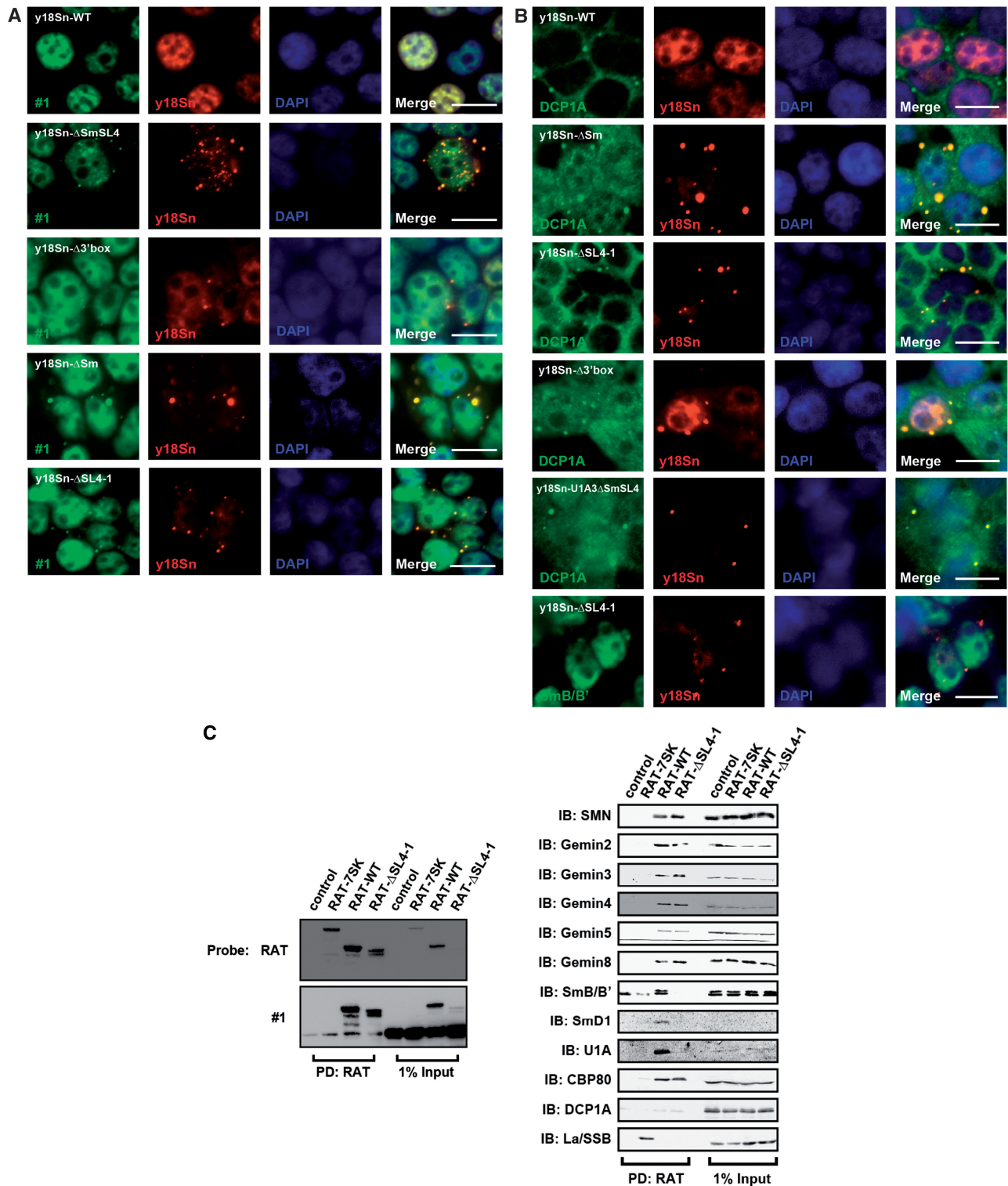


Figure 5. U1-tfs localize in P-bodies. (A) The 293T cells transfected with an expression vector encoding y18Sn-WT, y18Sn-ΔSmSL4, y18Sn-Δ3'box or y18Sn-ΔSm, or y18Sn-ΔSL4-1, were subjected to FISH. Endogenous U1 snRNA, exogenous U1 snRNA and U1-tfs were detected with probe #1 labeled with FITC (green). Exogenously expressed U1 snRNA or U1-tfs were also detected with the Cy3-labeled y18Sn probe (red). DAPI staining shows the nucleus. Merge: FITC, Cy3 and DAPI staining are merged. Scale bar: 10 μm. (B) y18Sn-WT-, y18Sn-ΔSm-, y18Sn-ΔSL4-1-, y18Sn-Δ3'box- or y18Sn-U1A3ΔSmSL4-expressing cells were stained by immunocytochemistry with antibodies against the proteins (green) indicated and by FISH with the Cy3-labeled y18Sn probe (red). (C) Proteins were pulled down (PD) from extract of RAT-WT- or RAT-ΔSL4-1-expressing cells or from control cells (co-transfected with vectors pcDNA3.1-PP7CP-HF and pcDNA3.1) by RAT-based affinity purification. RAT-tagged RNAs were detected by northern blotting with the RAT probe (left). PD: RAT, RAT-tagged RNA-protein complex bound to FLAG-tagged PP7CP was pulled down with anti-FLAG-conjugated beads and eluted with FLAG peptide. Proteins were visualized by immunoblotting (IB) with antibodies against the proteins indicated (right).

or y18Sn- Δ SL4-1-expressing cells, but fewer or no dot-stainings were observed in y18Sn-WT-expressing cells (Figure 5A). In y18Sn- Δ 3'box-expressing cells, few cytoplasmic dots were observed (Figure 5A and B). In y18Sn- Δ SmSL4- or y18Sn- Δ SL4-1-expressing cells, use of FITC-#1 or FITC-#3 yielded co-staining with Cy3-y18Sn in the cytoplasmic dots (Figure 5A and Supplementary Figure S5A). Those probes also showed nucleoplasmic staining (Figure 5 and Supplementary Figure S5A). On the other hand, FITC-#SL4 could not stain the Cy3-y18Sn-positive dots in y18Sn- Δ SL4-1-expressing cells (Supplementary Figure S5A), indicating that the molecular species lacking the SL4 region localized to the dots. In addition, these Cy3-y18Sn-stained dots co-localized with those stained with an antibody against SMN (Supplementary Figure S5B), Gemin2, Gemin5, Gemin6, Gemin8 (Supplementary Figure S5C) or DCP1A, the latter of which is a marker for P-bodies, in y18Sn- Δ SL4-1-expressing cells (Figure 5B). However, SmB/B' was not co-localized with the Cy3-y18Sn-stained dots in those cells (Figure 5B). On expression of y18Sn-WT, most of y18Sn-U1 snRNA was detected in the nucleus and its localization pattern was comparable to that of endogenous U1.

Using RAT-tagged RNAs as affinity bait, RNA-associated proteins were pulled down with anti-FLAG-conjugated beads from cell extracts of RAT-WT-, RAT- Δ SL4-1-, RAT- Δ 3'box-, RAT- Δ SL4-2- or RAT- Δ SL4-2 Δ 3'box-expressing cells co-expressing FLAG-tagged PP7CP and detected by immunoblotting. RAT-WT-expressing cells produced mostly RAT-U1 snRNA (Figure 5C), which associated with all the components of SMN complexes we examined (Figure 5C). We validated the specificity of RAT-tagged RNA by using RAT-7SK, which associated with La/SSB as reported (Figure 5C) (36). On the other hand, RAT- Δ SL4-1-expressing cells produced mainly RAT-U1-tfs (Figure 5C). The pull-down showed that RAT-U1-tfs associated with all the proteins that bind RAT-U1 snRNA except SmB/B', SmD1, and U1A (Figure 5C). Similar results were obtained by using RAT- Δ SL4-2 and RAT- Δ SL4-2 Δ 3'box-expressing cells (Supplementary Figure S6A). RAT- Δ 3'box-expressing cells produced not only RAT-U1 snRNA but also U1-tfs; accordingly, SmB/B' was pulled down less in comparison with the other proteins pulled down by RAT-U1 snRNA or RAT-U1-tfs (Supplementary Figure S6A). These results indicated that U1-tfs forms a complex with most of the known components of SMN complexes except SmB/B', SmD1 and U1A and suggested that U1-tfs are transported to P-bodies with those proteins.

To validate RNA-tagged RNA species in terms of post-transcriptional modifications, we analyzed the pulled-down RAT-U1 snRNA (produced from RAT-WT) and RAT-U1-tfs (produced from RAT- Δ SL4-1 or RAT- Δ SL4-1 Δ 3'box) using the LC-MS/MS-Ariadne method after RNase T1 digestion of their corresponding SYBR Gold-stained RNA bands excised from urea-PAGE gels. RAT-WT produced U1 having the m³G cap (~99% of the U1 population) and methylation at position 70 (~40% of the population), but RAT-U1-tfs had the

m⁷G cap exclusively and had base methylation at the first adenine in ~50% of its population (Supplementary Figure S6B and Supplementary Table S4). Thus, RAT-tagged transcripts underwent post-transcriptional modifications in a manner similar to that observed for the corresponding endogenous transcripts.

U1-tfs are degraded more rapidly than mature-U1 snRNA

Given the report that P-bodies function in RNA surveillance and turnover (30), we compared the degradation rate of U1-tfs with that of mature U1 snRNA. We transfected y18Sn-WT, y18Sn- Δ SmSL4 or y18Sn- Δ 3' box construct in 293T cells, treated the cells with actinomycin D and measured the time-dependent changes in the cellular levels of U1 snRNA and U1-tfs in the total RNA. As clearly shown in Figure 6A, the level of U1-tfs decreased much more rapidly than that of mature U1 snRNA, suggesting that U1-tfs are degraded efficiently in P-bodies. Given that U1-tfs co-localized with SMN and its associated components in P-bodies, we next addressed whether SMN plays an active role in this degradation pathway, such that SMN participates in P-body localization and degradation of U1-tfs. To examine this, we took advantage of previously characterized mutations in the SL1 region of U1 snRNA (U1A3) that abolish U1 binding to SMN, as reported by Yong *et al.* (11). We constructed the y18Sn-U1A3 Δ SmSL4 vector, expressed the construct in 293T cells and detected y18Sn-tagged RNA by northern blot or *in situ* hybridization analysis. If SMN participates in the P-body localization and degradation of U1-tfs, this transcript will not be localized into P-bodies and will not be degraded efficiently. Those analyses, however, detected only a minute amount of the truncated form of y18Sn-U1A3 Δ SmSL4 (Figure 6B) but showed dominant P-body localization of y18Sn-U1A3 Δ SmSL4 (Figure 5B). These results suggested that SMN did not have a direct role in the P-body localization of U1-tfs, although we could not evaluate the role of SMN in the degradation rate of U1-tfs because of its extremely low expression level in the cells.

SmB/B' deficiency or deletion of the Sm-binding site induces P-body localization of U1 snRNP

Given our results that U1-tfs are localized in P-bodies but do not associate with SmB/B' or SmD1, we considered the possibility that P-body localization of U1-tfs is induced by failure of the assembly of Sm proteins on the Sm site. We, therefore, postulated that SmB/B' deficiency would cause full-length U1 snRNA to localize to P-bodies. Knockdown of SmB/B' with stealth RNA (siRNA1; si-1) increased the number of P-bodies in the cytoplasm compared with a control RNA (scrambled sequence), as detected with an antibody against DEAD (Asp-Glu-Ala-Asp) box protein 6 (DDX6), Sm-like protein (LSM)14A, Gemin5, or DCP1A, which is a marker for P-bodies (Figure 7A–C). SmB/B' knockdown reduced the level of U1 snRNA (Supplementary Figure S7A) and induced P-body localization of not only endogenous U1 snRNA but also SMN and Gemin5 (Figure 7A–C). Use of another siRNA (siRNA2; si-2 that has a different sequence from si-1) for SmB/B' knockdown also showed P-body

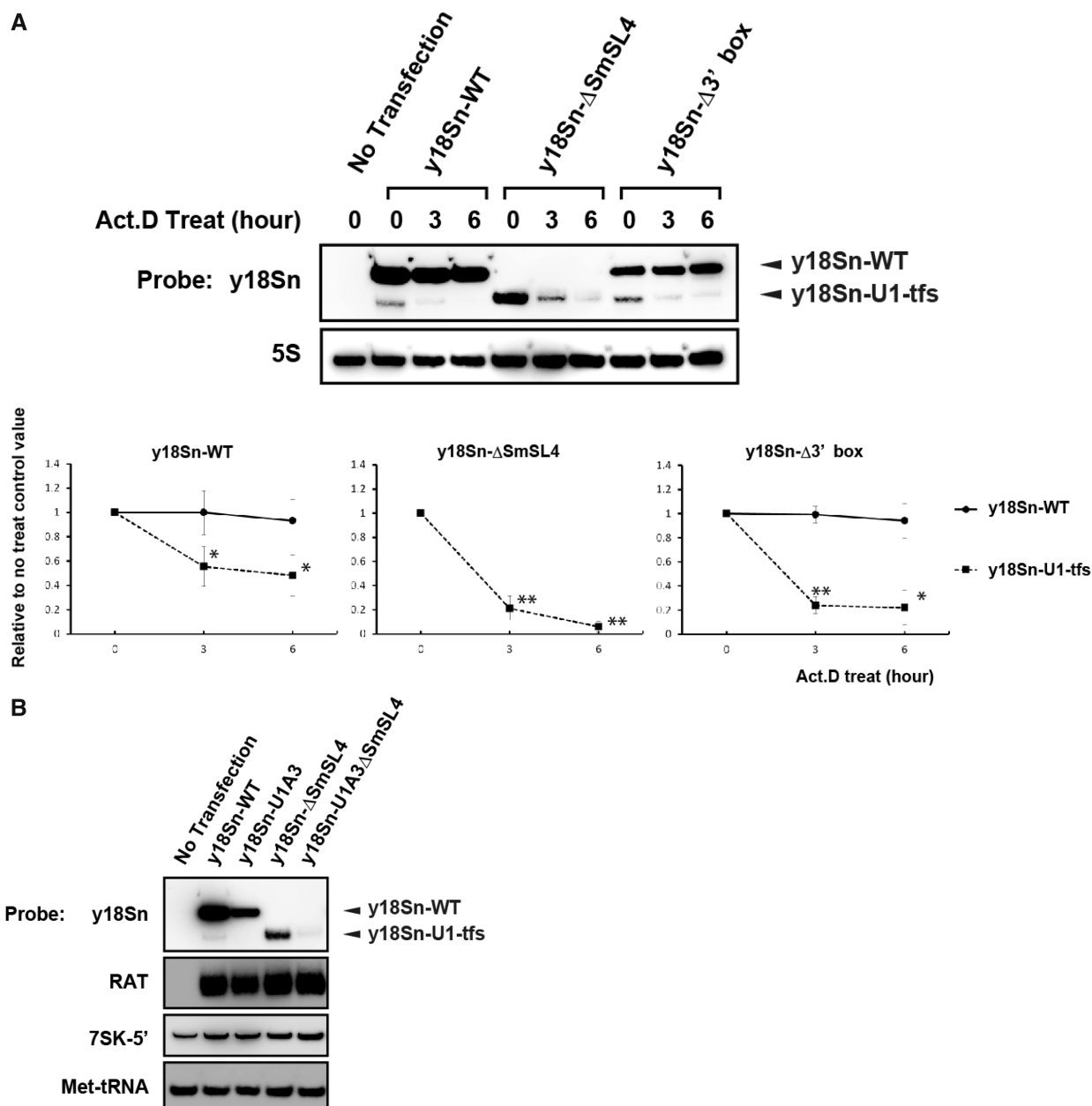


Figure 6. U1-tfs are degraded more rapidly than mature-U1 snRNA is. (A) RNAs were prepared from y18Sn-WT-, y18Sn-ΔSmSL4- or y18SnΔ3' box-expressing 293T cells at 0, 3 or 6 h after 1 μg/ml actinomycin D (Act.D) treatment, and analyzed by northern blotting with the probe for y18Sn or 5S rRNA (5S) (top). (Bottom) The y-axis shows the staining intensities relative to that of the cells harvested at 0 h. The values indicated are averages (±SD) of three independent experiments. * $P < 0.05$. ** $P < 0.01$. (B) RNAs were prepared from y18Sn-WT-, y18Sn-U1A3-, y18Sn-ΔSmSL4- or y18Sn-U1A3ΔSmSL4-expressing 293T cells and analyzed by northern blotting with the probes indicated. RAT-tag vector was used as a transfection control.

localization of U1 snRNA (Supplementary Figure S7B). We also observed that the knockdown with si-1 and si-2 reduced the level of U2 snRNA and induced P-body localization of U2 snRNA, respectively (Supplementary Figure S7A, S7C and S7D). SmB/B'-deficient cells also showed reduced staining for FITC-#1 and Cy3-U2 (Figure 7B and Supplementary Figure S7C and D), consistent with reduced SYBR Gold staining for U1 and U2 snRNA in total RNA (Supplementary Figure S7A). Furthermore, y18SnΔSm RNA lacking the Sm-binding site but not the SL4 region, co-localized with SMN in

P-bodies (Supplementary Figure S5B). However, the siRNA-mediated knockdown of SMN did not cause P-body localization of y18Sn-WT that was co-transfected with the siRNA (Supplementary Figure S7E). These results strongly suggested that the failure of Sm protein loading to U1 snRNA-induced localization of the U1 snRNA-SMN complex to P-bodies and accelerated the degradation of U1 snRNA (Figure 7A–C and Supplementary Figure S7A), whereas SMN did not have a direct role in the P-body localization of U1 snRNA and U1-tfs (Supplementary Figure S7E).

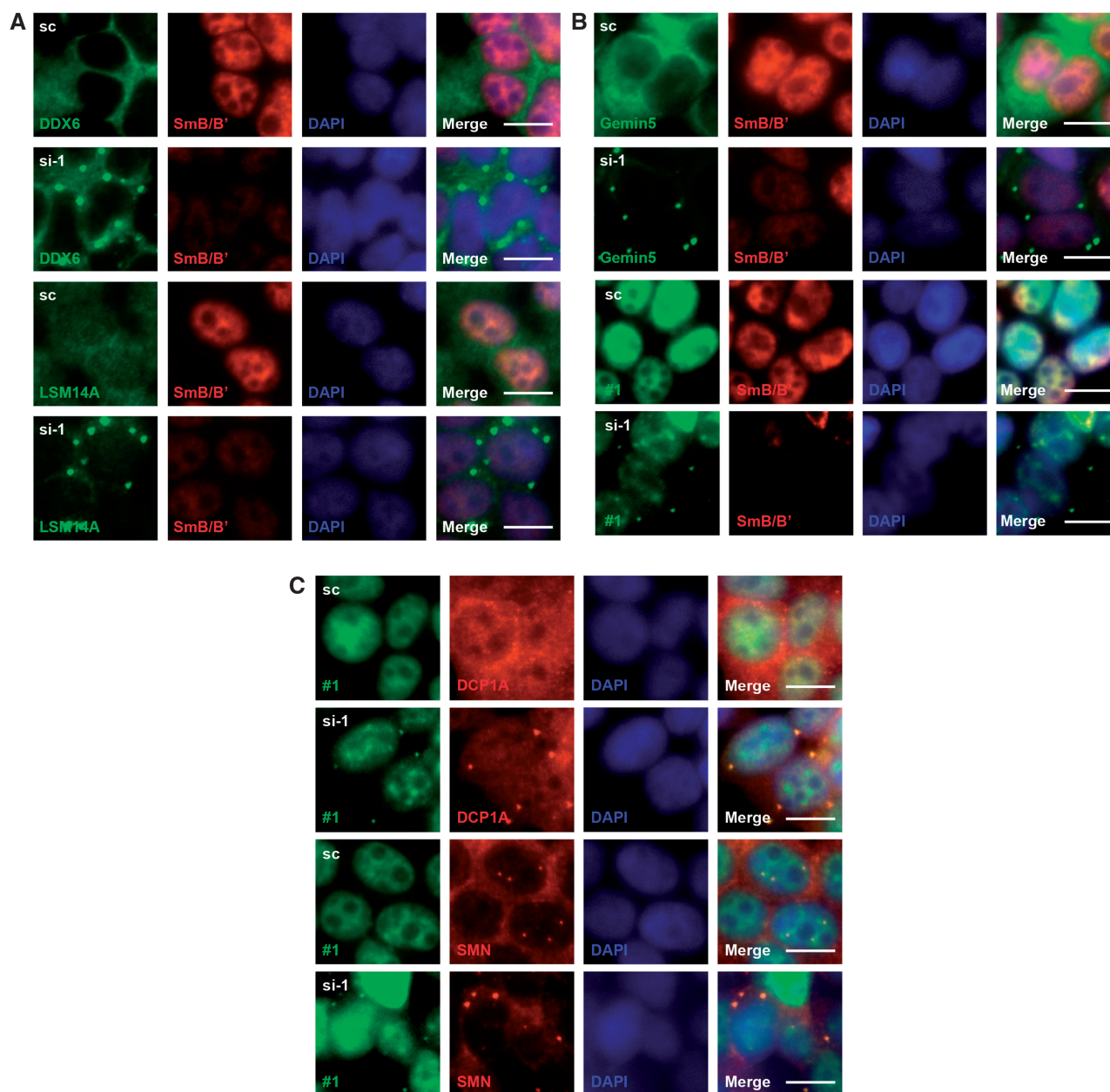


Figure 7. Failure of SmB/B' loading increases the number of P-bodies and causes P-body localization of U1 snRNA. (A) 293T cells transfected transiently with a stealth SmB/B' siRNA (si-1) or scrambled RNA (sc) were analyzed by immunocytochemistry using anti-SmB/B' primary antibody and Cy3-labeled secondary antibody (red). The other proteins were detected with their corresponding antibodies and FITC-labeled secondary antibody (green). DAPI staining shows the nucleus. Merge: FITC, Cy3 and DAPI staining are merged. Scale bar: 10 μ m. (B) SmB/B' was detected with anti-SmB/B' primary antibody and Cy3-labeled secondary antibody (red). Gemin5 was detected with its antibody (green). U1 snRNA was detected by FISH with probe #1 (green). (C) Cells were analyzed by combination of FISH with probe #1 (green) and immunocytochemistry with the antibodies indicated (red).

DISCUSSION

In this study, we demonstrated that U1-tfs were formed at steps before 5' cap hypermethylation, were diverted from the canonical pathway of U1 snRNP biogenesis owing to the failure of Sm protein loading onto U1-tfs lacking the Sm-binding site, and were destined for localization in P-bodies where surveillance and degradation of RNA take place (Figure 8). We uncovered at least an additional pathway to the degradation of aberrant snRNAs in addition to a known quality control step in snRNP biogenesis.

On retrieval of sequences for the U1 genes or related genes from the genome database reference assembly (version GRCh37), we localized 41 loci for these genes in the human genome (Supplementary Table S5). Of these, 3 loci contain the U1 snRNA coding region and all 3 cis-acting elements (DSE, PSE and 3' box) (full U1 genes), 7 loci contain DSE and PSE but no 3' box (Δ 3' box-U1 genes), 4 loci contain only DSE (DSE-U1 genes), 5 loci contain only PSE (PSE-U1 genes) and the remaining 22 loci have the U1 coding region but lack the *cis*-acting

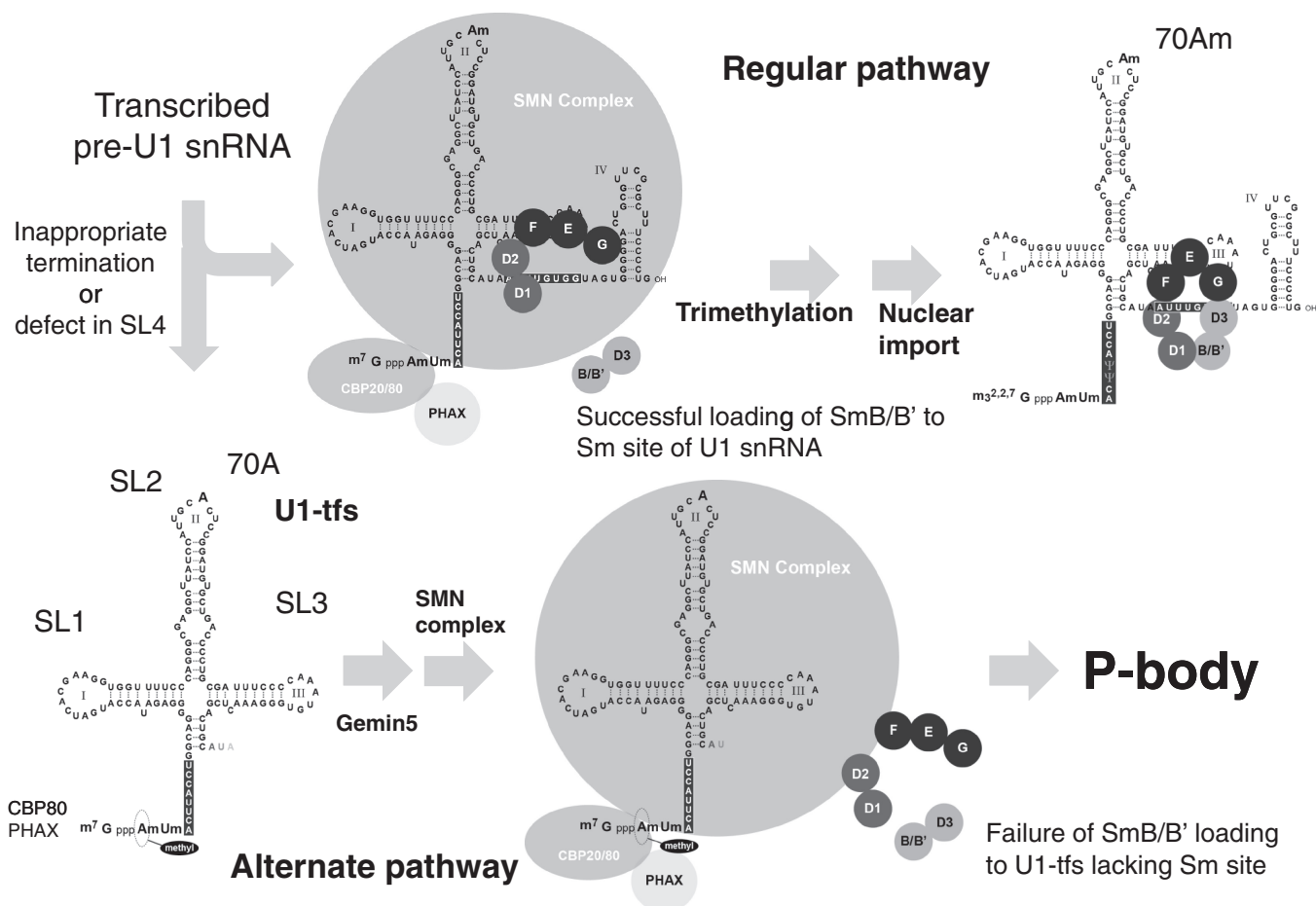


Figure 8. Proposed alternate pathway for U1 snRNP biogenesis. In an early cytoplasmic stage of the canonical pathway of U1 snRNP biogenesis, the SMN complex containing Gemin 2–8 brings the pentamer of Sm proteins (SmD1, D2, E, F and G) and pre-U1 snRNA together, and then SmB/B' and D3 are incorporated into the Sm pentamer to form a heptameric Sm ring. Successful formation of the Sm heptamer around the Sm site of pre-U1 snRNA leads to trimethylation of the 5' cap of pre-U1 snRNA followed by nuclear import to form the mature U1 snRNP with ribose methylation at position 70 (adenosine; 70 Am) of U1 snRNA. In the alternate pathway at an early stage of U1 snRNP biogenesis, some pre-U1 snRNAs with inappropriate 3' termination or defects in the SL4 region are transformed to U1-tfs with a m⁷G cap; U1-tfs lack the Sm site and SL4 region and are methylated at position 1 (adenosine) but not at the position-70 ribose. Because U1-tfs lack the Sm site, they are unable to form the Sm heptamer owing to its inability to accept SmB/B' (and probably SmD3) loading to the Sm pentamer on the SMN complex. Failure of SmB/B' (and probably SmD3) loading and of formation of the heptameric Sm ring on the pre-snRNA results in an increased number of P-bodies and/or localization of the U1-tfs–SMN complex to P-bodies, which function in RNA surveillance and decay.

elements (U1 pseudogenes). Our present results indicated that not only full-length U1 genes but also $\Delta 3'$ box-U1 and PSE-U1 genes could be transcribed and produce mature U1 snRNA, suggesting that full-length U1 genes and $\Delta 3'$ box-U1 genes contribute to the formation of U1-tfs. Given that the 3' box element is required for appropriate processing with the integrator complex at the 3' end of pre-U1 snRNA (46), we propose that failure of 3' end processing or inappropriate processing leads to the formation of U1-tfs. When transcription reaches the 3' box cis-acting element, the integrator complex cleaves pre-U1 snRNA at the proper 3' site. Cleaved pre-U1 snRNA is then trimmed by a nuclear 3'–5' exonuclease. Failure or inefficiency in cleavage of pre-U1 snRNA by the integrator complex during transcription of $\Delta 3'$ box-U1 genes and less frequently during that of full-length U1 genes may result in a pre-U1 snRNA longer than that

transcribed from full-length U1 genes, which may perturb the secondary structure of the 3' end region of the pre-U1 snRNA. In addition, defects in the SL4 region of U1 genes may result in the formation of U1-tfs, indicating the importance of the SL4 region of pre-U1 snRNA for appropriate formation of mature U1 snRNA; therefore, we also suggest that mutation in or aberrant transcription of the SL4 region leads to the formation of U1-tfs.

The analyses of Gemin5-associated complexes prepared from the fractions separated by ultracentrifugation suggested that PHAX- and Gemin5-associated U1-tfs were formed first and then were combined with the SMN complex composed of Gemin 2, 8 and unrip (Figure 3B and C). Pull-down analysis using RAT-U1-tfs (Figure 5C) also showed that the U1-tfs–SMN complex contained almost all the components of the known SMN complex

except SmB/B', SmD1 and U1A. Failure to incorporate Sm proteins into the Sm site perhaps results in the inability to displace the N-terminal region of Gemin2, and therefore the heptameric Sm ring cannot be fully formed. The U1-tfs-SMN complex localized almost exclusively in P-bodies (Figure 5A and B and Supplementary Figure S5B), an aspect distinct from the normal U1 snRNA-SMN complex (Figure 5A and B and Supplementary Figure S5B); thus, the complex was discriminated from the normal U1 snRNA-SMN complex in terms of cellular localization. Use of y18Sn-U1A3ΔSmSL4 mutant that lacked the ability to bind to SMN, however, suggested that SMN had no active role in this discrimination (Figure 5B).

Because the U1-tfs-SMN complex lacks at least SmB/B' and SmD1, we hypothesized that this discrimination is a consequence of the failure of Sm protein loading into the Sm pentamer's RNA-binding pocket. This hypothesis was substantiated by the result that SmB/B' knockdown or the expression of a truncated U1 snRNA lacking the Sm-binding site (y18Sn-ΔSm) increased the number of P-bodies (Figure 5B); moreover, SmB/B' knockdown induced P-body localization of endogenous U1 snRNA (Figure 7B). Based on these data coupled with the result that U1-tfs were degraded more rapidly than was mature-U1 snRNA (Figure 6A), we conclude that a novel surveillance pathway for U1 snRNA exists in human cells; namely, inappropriate termination or defects in the SL4 region of U1 snRNA cause processing to U1-tfs lacking the Sm site, which is then eliminated through P-bodies (Figure 8). The results that SmB/B' knockdown reduced the cellular level of U1 snRNA (Figure 7B and Supplementary Figure S7A) also support this conclusion.

We also observed that SmB/B' knockdown reduced the cellular level of U2 snRNA (Supplementary Figure S7A). Similar results were reported by Saltzman *et al.* (47), in which depletion of SmB/B' in human cells resulted in reduced levels of Sm-class snRNAs (U1, U4, U5, U11, U12 and U4atac) but not LSm-class snRNAs (U6 and U6atac), and similar reduction of U2 snRNA was observed on knockdown of SmD1 (47). These reports as well as our present results suggest that a general snRNA surveillance mechanism may facilitate the elimination of inappropriate Sm-class snRNAs via the recognition of truncated forms of those snRNAs in SMN complexes. It will thus be intriguing to investigate whether short forms of other Sm-class snRNAs exist and whether they are eliminated via a similar pathway.

Our present study raises many questions, one of which concerns how U1-tfs are formed. Are U1-tfs formed in the nucleus or after transport to the cytoplasm? Which ribonuclease is involved? What is the mechanism by which U1-tfs transcripts are distinguished from normal U1 transcripts? Do differences in post-transcriptional modifications at positions 1 and 70 between U1-tfs and mature U1 snRNA contribute to this discrimination mechanism? Addressing these issues will require a fuller understanding of the surveillance pathway of U snRNAs.

Recently, apart from its roles in splicing, U1 snRNP was reported to regulate transcript length through co-transcriptional recognition of cryptic polyadenylation

signals and inhibition of premature cleavage and polyadenylation at these sites (48,49). Thus, the cellular level of U1 snRNP relative to the level of general transcription may determine the transcriptome and proteome of specific cell types. Our present results raise the additional possibility that the level of U1 snRNP is regulated by cellular levels of proteins, such as SmB/B', for which an absence invokes the surveillance pathway we have identified.

Our present study provides a good example of the beneficial use of MS for discovering a new degradation pathway during U1 snRNP biogenesis. In general, RNPs function in many cellular processes. Elucidating the cooperative actions between RNAs and proteins is important for understanding biological systems. Post-transcriptional or metabolic modifications of RNA play vital roles in cooperative actions with proteins. Direct analysis of RNA using MS allows unbiased identification of RNA and has great advantages in high-throughput analysis to acquire information about RNA modifications (31–34). In many diseases, especially neurological diseases, RNA metabolism is altered, and thus the methods for direct analysis of RNAs described in our present study may be useful for understanding the pathogenesis of those diseases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Dr. J. Robert Hogg and Dr. Kathleen Collins provided pET28ZZTPP7His and pRcU6PT7SK plasmid vectors, respectively, for RAT purification. Dr. Robin Reed provided anti-CBP80. Dr. Gideon Dreyfuss provided anti-SMN (2B1), Gemin2 (2E17), Gemin3 (12H12), Gemin4 (17D10), Gemin6 (20H8), Gemin8 (1F8) and Unrip (3G6). The authors also thank Dr. Gideon Dreyfuss and Dr. Lili Wan for critical reading of this manuscript and for their valuable suggestions and comments.

FUNDING

Grant for Core Research for Evolutionary Science and Technology (CREST) from Japan Science and Technology Agency [grant ID 13415564]. Grant-in-Aid for Scientific Research from The Ministry of Education, Culture, Sports, Science, & Technology of Japan (MEXT) [grant No. 24241075]. Funding for open access charge: Core Research for Evolutionary Science and Technology (CREST), Japan Science and Technology Agency.

Conflict of interest statement. None declared.

REFERENCES

- Patel, S.B. and Bellini, M. (2008) The assembly of a spliceosomal small nuclear ribonucleoprotein particle. *Nucleic Acids Res.*, **36**, 6482–6493.

2. Mattaj, I.W. (1986) Cap trimethylation of U snRNA is cytoplasmic and dependent on U snRNP protein binding. *Cell*, **46**, 905–911.
3. Meister, G., Eggert, C. and Fischer, U. (2002) SMN-mediated assembly of RNPs: a complex story. *Trends Cell Biol.*, **12**, 472–478.
4. Fischer, U., Liu, Q. and Dreyfuss, G. (1997) The SMN-SIP1 complex has an essential role in spliceosomal snRNP biogenesis. *Cell*, **90**, 1023–1029.
5. Yong, J., Kasim, M., Bachorik, J.L., Wan, L. and Dreyfuss, G. (2010) Gemin5 delivers snRNA precursors to the SMN complex for snRNP biogenesis. *Mol. Cell*, **38**, 551–562.
6. Wahl, M.C., Will, C.L. and Lührmann, R. (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell*, **136**, 701–718.
7. Lunn, M.R. and Wang, C.H. (2008) Spinal muscular atrophy. *Lancet*, **371**, 2120–2133.
8. Workman, E., Kolb, S.J. and Battle, D.J. (2012) Spliceosomal small nuclear ribonucleoprotein biogenesis defects and motor neuron selectivity in spinal muscular atrophy. *Brain Res.*, **1462**, 93–99.
9. Burghes, A.H. and Beattie, C.E. (2009) Spinal muscular atrophy: why do low levels of survival motor neuron protein make motor neurons sick? *Nat. Rev. Neurosci.*, **10**, 597–609.
10. Pellizzoni, L., Yong, J. and Dreyfuss, G. (2002) Essential role for the SMN complex in the specificity of snRNP assembly. *Science*, **298**, 1775–1779.
11. Yong, J., Pellizzoni, L. and Dreyfuss, G. (2002) Sequence-specific interaction of U1 snRNA with the SMN complex. *EMBO J.*, **21**, 1188–1196.
12. Yong, J., Wan, L. and Dreyfuss, G. (2004) Why do cells need an assembly machine for RNA-protein complexes? *Trends Cell Biol.*, **14**, 226–232.
13. Lau, C.K., Bachorik, J.L. and Dreyfuss, G. (2009) Gemin5-snRNA interaction reveals an RNA binding function for WD repeat domains. *Nat. Struct. Mol. Biol.*, **16**, 486–491.
14. Zhang, R., So, B.R., Li, P., Yong, J., Glisovic, T., Wan, L. and Dreyfuss, G. (2011) Structure of a key intermediate of the SMN complex reveals Gemin2's crucial function in snRNP assembly. *Cell*, **146**, 384–395.
15. Grimm, C., Chari, A., Pelz, J.P., Kuper, J., Kisker, C., Diederichs, K., Stark, H., Schindelin, H. and Fischer, U. (2013) Structural basis of assembly chaperone-mediated snRNP formation. *Mol. Cell*, **49**, 692–703.
16. Sarachan, K.L., Valentine, K.G., Gupta, K., Moorman, V.R., Gledhill, J.M. Jr, Bernens, M., Tommos, C., Wand, A.J. and Van Duyne, G.D. (2012) Solution structure of the core SMN-Gemin2 complex. *Biochem. J.*, **445**, 361–370.
17. Neuenkirchen, N., Chari, A. and Fischer, U. (2008) Deciphering the assembly pathway of Sm-class U snRNPs. *FEBS Lett.*, **582**, 1997–2003.
18. Pellizzoni, L., Charroux, B. and Dreyfuss, G. (1999) SMN mutants of spinal muscular atrophy patients are defective in binding to snRNP proteins. *Proc. Natl Acad. Sci. USA*, **96**, 11167–11172.
19. Brahms, H., Meheus, L., de Brabandere, V., Fischer, U. and Lührmann, R. (2001) Symmetrical dimethylation of arginine residues in spliceosomal Sm protein B/B' and the Sm-like protein LSm4, and their interaction with the SMN protein. *RNA*, **7**, 1531–1542.
20. Sprangers, R., Groves, M.R., Sinning, I. and Sattler, M. (2003) High-resolution X-ray and NMR structures of the SMN Tudor domain: conformational variation in the binding site for symmetrically dimethylated arginine residues. *J. Mol. Biol.*, **327**, 507–520.
21. Friesen, W.J., Massenet, S., Paushkin, S., Wyce, A. and Dreyfuss, G. (2001) SMN, the product of the spinal muscular atrophy gene, binds preferentially to dimethylarginine-containing protein targets. *Mol. Cell*, **7**, 1111–1117.
22. Friesen, W.J., Paushkin, S., Wyce, A., Massenet, S., Pesiridis, G.S., Van Duyne, G., Rappsilber, J., Mann, M. and Dreyfuss, G. (2001) The methylosome, a 20S complex containing JBP1 and pICln, produces dimethylarginine-modified Sm proteins. *Mol. Cell Biol.*, **21**, 8289–8300.
23. Meister, G., Eggert, C., Bühler, D., Brahms, H., Kambach, C. and Fischer, U. (2001) Methylation of Sm proteins by a complex containing PRMT5 and the putative U snRNP assembly factor pICln. *Curr. Biol.*, **11**, 1990–1994.
24. Carissimi, C., Saieva, L., Baccon, J., Chiarella, P., Maiolica, A., Sawyer, A., Rappsilber, J. and Pellizzoni, L. (2006) Gemin8 is a novel component of the survival motor neuron complex and functions in small nuclear ribonucleoprotein assembly. *J. Biol. Chem.*, **281**, 8126–8134.
25. Ma, Y., Dostie, J., Dreyfuss, G. and Van Duyne, G.D. (2005) The Gemin6-Gemin7 heterodimer from the survival of motor neurons complex has an Sm protein-like structure. *Structure*, **13**, 883–892.
26. Ogawa, C., Usui, K., Ito, F., Itoh, M., Hayashizaki, Y. and Suzuki, H. (2009) Role of survival motor neuron complex components in small nuclear ribonucleoprotein assembly. *J. Biol. Chem.*, **284**, 14609–14617.
27. Carissimi, C., Saieva, L., Gabanella, F. and Pellizzoni, L. (2006) Gemin8 is required for the architecture and function of the survival motor neuron complex. *J. Biol. Chem.*, **281**, 37009–37016.
28. Otter, S., Grimm, M., Neuenkirchen, N., Chari, A., Sickmann, A. and Fischer, U. (2007) A comprehensive interaction map of the human survival of motor neuron (SMN) complex. *J. Biol. Chem.*, **282**, 5825–5833.
29. Liu, J.L. and Gall, J.G. (2007) U bodies are cytoplasmic structures that contain uridine-rich small nuclear ribonucleoproteins and associate with P bodies. *Proc. Natl Acad. Sci. USA*, **104**, 11655–11659.
30. Lee, L., Davies, S.E. and Liu, J.L. (2009) The spinal muscular atrophy protein SMN affects Drosophila germline nuclear organization through the U body-P body pathway. *Dev. Biol.*, **332**, 142–155.
31. Nakayama, H., Akiyama, M., Taoka, M., Yamauchi, Y., Nobe, Y., Ishikawa, H., Takahashi, N. and Isobe, T. (2009) Ariadne: a database search engine for identification and chemical analysis of RNA using tandem mass spectrometry data. *Nucleic Acids Res.*, **37**, e47.
32. Taoka, M., Yamauchi, Y., Nobe, Y., Masaki, S., Nakayama, H., Ishikawa, H., Takahashi, N. and Isobe, T. (2009) An analytical platform for mass spectrometry-based identification and chemical analysis of RNA in ribonucleoprotein complexes. *Nucleic Acids Res.*, **37**, e140.
33. Taoka, M., Ikumi, M., Nakayama, H., Masaki, S., Matsuda, R., Nobe, Y., Yamauchi, Y., Takeda, J., Takahashi, N. and Isobe, T. (2010) In-gel digestion for mass spectrometric characterization of RNA from fluorescently stained polyacrylamide gels. *Anal. Chem.*, **82**, 7795–7803.
34. Nakayama, H., Takahashi, N. and Isobe, T. (2011) Informatics for mass spectrometry-based RNA analysis. *Mass Spectrom. Rev.*, **30**, 1000–1012.
35. Pomeranz Krummel, D.A., Oubridge, C., Leung, A.K., Li, J. and Nagai, K. (2009) Crystal structure of human spliceosomal U1 snRNP at 5.5 Å resolution. *Nature*, **458**, 475–480.
36. Hogg, J.R. and Collins, K. (2007) RNA-based affinity purification reveals 7SK RNPs with distinct composition and regulation. *RNA*, **13**, 868–880.
37. Rouquette, J., Choesmel, V. and Gleizes, P.E. (2005) Nuclear export and cytoplasmic processing of precursors to the 40S ribosomal subunits in mammalian cells. *EMBO J.*, **24**, 2862–2872.
38. Hayano, T., Yamauchi, Y., Asano, K., Tsujimura, T., Hashimoto, S., Isobe, T. and Takahashi, N. (2008) Automated SPR-LC-MS/MS system for protein interaction analysis. *J. Proteome Res.*, **7**, 4183–4190.
39. Paushkin, S., Gubitz, A.K., Massenet, S. and Dreyfuss, G. (2002) The SMN complex, an assemblyosome of ribonucleoproteins. *Curr. Opin. Cell Biol.*, **14**, 305–312.
40. Hebert, M.D., Szymczyk, P.W., Shpargel, K.B. and Matera, A.G. (2001) Coilin forms the bridge between Cajal bodies and SMN, the spinal muscular atrophy protein. *Genes Dev.*, **15**, 2720–2729.
41. Massenet, S., Mougin, A. and Branlant, C. (1998) Posttranscriptional modifications in the U small nuclear RNAs. In: Grosjean, H. and Benne, R. (eds), *The Modification and Editing of RNA*. ASM Press, Washington, DC, pp. 201–227.
42. Ohno, M., Segref, A., Bachi, A., Wilm, M. and Mattaj, I.W. (2000) PHAX, a mediator of U snRNA nuclear export whose activity is regulated by phosphorylation. *Cell*, **101**, 187–198.

43. Fortes,P., Bilbao-Cortés,D., Fornerod,M., Rigaut,G., Raymond,W., Séraphin,B. and Mattaj,I.W. (1999) Luc7p, a novel yeast U1 snRNP protein with a role in 5' splice site recognition. *Genes Dev.*, **13**, 2425–2438.
44. Egloff,S., O'Reilly,D. and Murphy,S. (2008) Expression of human snRNA genes from beginning to end. *Biochem. Soc. Trans.*, **36**, 590–594.
45. Fujii,K., Kitabatake,M., Sakata,T., Miyata,A. and Ohno,M. (2009) A role for ubiquitin in the clearance of nonfunctional rRNAs. *Genes Dev.*, **23**, 963–974.
46. Baillat,D., Hakimi,M.A., Nääär,A.M., Shilatifard,A., Cooch,N. and Shiekhattar,R. (2005) Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell*, **123**, 265–276.
47. Saltzman,A.L., Pan,Q. and Blencowe,B.J. (2011) Regulation of alternative splicing by the core spliceosomal machinery. *Genes Dev.*, **25**, 373–384.
48. Kaida,D., Berg,M.G., Younis,I., Kasim,M., Singh,L.N., Wan,L. and Dreyfuss,G. (2010) U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, **468**, 664–668.
49. Berg,M.G., Singh,L.N., Younis,I., Liu,Q., Pinto,A.M., Kaida,D., Zhang,Z., Cho,S., Sherrill-Mix,S., Wan,L. *et al.* (2012) U1 snRNP determines mRNA length and regulates isoform expression. *Cell*, **150**, 53–64.
50. Sambrook,J. and Russell,D. (2001) *Molecular Cloning: A Laboratory Manual*, 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
51. Natsume,T., Yamauchi,Y., Nakayama,H., Shinkawa,T., Yanagida,M., Takahashi,N. and Isobe,T. (2002) A direct nanoflow liquid chromatography-tandem mass spectrometry system for interaction proteomics. *Anal. Chem.*, **74**, 4725–4733.
52. McLuckey,S.A., Van Berkel,G.J. and Glish,G.L. (1992) Tandem mass spectrometry of small multiply charged oligonucleotides. *J. Am. Soc. Mass Spectrom.*, **3**, 60–70.